DOCUMENT RESUME

ED 340 778 TM 018 034

AUTHOR Linn, Robert L.
TITLE Test Misuse: Why Is It So Prevalent?
SPONS AGENCY Congress of the U.S., Washington, D.C. Office of Technology Assessment.
PUB DATE Sep 91
NOTE 11p.; Contractor report prepared for the Office of Technology Assessment titled "Testing in American Schools: Asking the Right Questions." For related document, see TM 018 025.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Communication Skills; Educational Assessment; Educational History; Educational Technology; Elementary Secondary Education; *Evaluation Utilization; Information Dissemination; *Political Influences; *Scores; *Standardized Tests; Student Evaluation; Test Coaching; *Testing Problems; Test Interpretation; Test Results; Test Use

ABSTRACT

Issues in the misuse of standardized test scores are discussed. While information misuse is not a problem unique to testing, it is particularly prevalent. Information can be a source of power, and politically motivated uses of information explain part of the misuse of test scores. Other misuses have historical roots in the exaggerated claims of early testers for their new technology. Popular interpretations often make the mistake of assuming that the reason for a poor performance can be inferred from the score. Test results depend on a host of contextual factors of test use and administration, many of which are not known to the public. The Lake Wobegon effect (the tendency of states and districts to report scores above the national average) is largely the result of the reuse of tests and of changes in the stakes attached to test results for teachers and the school. The higher the stakes, the more likely test preparation is to be extensive. It is not reasonable to expect the public and the media to become testing experts, but it is appropriate for test specialists to become more sensitive to issues of correct reporting and test interpretation. Communicating test results to a wider audience requires more attention than it has generally received. Recently, there was a concerted effort to encourage proper interpretations and discourage improper ones on the part of the National Assessment Governing Board (NAGB), which has policy oversight of National Assessment of Educational Progress (NAEP); the sponsoring governmental agency, the National Center of Educational Statistics (NCES); and the primary contractor, Educational Testing Service (ETS). (SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED340728

# TEST MISUSE: WHY IS IT SO PREVALENT?

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

by
Robert L. Linn
Center for Research on Evaluation, Standards, and Student Testing
University of Colorado at Boulder

for
Office of Technology Assessment
U.S. Congress

September 1991

TM018034

BEST COPY AVAILABLE

**Test Misuse: Why Is It So Prevalent?**

Robert L. Linn

Center for Research on Evaluation, Standards, and Student Testing, University of Colorado at Boulder

Standardized test scores are used for a host of purposes other than those for which they were originally designed. All too frequently, these expanded uses are misuses that sometimes lead to serious misinterpretations of educational achievement. An achievement test that is intended to help teachers identify student strengths and weaknesses, for example, may end up being misused to hold teachers accountable for student achievement, to rank schools within a district, or even to help sell real estate. Similarly, the Scholastic Aptitude Test (SAT), which is designed to "supplement the high school record and other information in assessing a student's competence for college work",[1] is regularly misused, to rank states and interpreted as if the results indicated the relative quality of education being provided by different states. The fact that the SAT is not even an achievement test, much less a measure of the quality of an educational system, and the fact that the rank order of states is highly predictable from the proportion of students within a state who take the test, are simply ignored.

The list of misuses could easily be expanded, indeed, many discussions of misuses of test results already exist. In any event, the more interesting and vexing question is why are the misuses so prevalent? Although, at some level of analysis, there

may be nearly as many answers to this question as there are common misuses, there are also a few likely culprits that may explain a large fraction of the misuses.

Clearly, information misuse is not a problem that is unique to test results. Information can be a source of power, a means of buttressing one's own political agenda or attacking an opposing position. Crime statistics, economic indicators, and many other types of quantitative information are regularly twisted, partially reported, and conveniently interpreted in ways that best serve particular policies or political positions.

A political candidate who wants to run on an educational reform platform finds that low achievement test scores provide a powerful means of making the case that the state's schools are in dire straights. A school superintendent points with pride to rising district test scores as evidence of improvement, while ignoring their possibly misleading nature due continued reuse of the same test or possibly negative side effects due to narrowing the focus of instruction to preparation for that particular test. Politically motivated uses of information surely explain a part of the misuse of test results. But the problem of test misuse is much greater than that; many of the misuses require other explanations.

Some misuses have deep historical roots. Early testers were overly enthusiastic about their new technology. They made exaggerated claims about what tests could measure and about what could be accomplished with them. They wrapped their claims in a scientific aura. The mistaken notion that tests could yield

direct measures of genetic capacity and the associated misinterpretations of ethnic and racial group differences left a sad legacy.

Contemporary leaders in educational measurement have a much more modest view of what tests measure and how they should be interpreted than was common among early testers. Unfortunately, the popul - view of tests is too frequently more in tune with the exaggerated claims of early testers than with the more modest views of contemporary specialists in educational measurement.

A limitation of any test that is obvious to testing specialists is that a test score can only describe a level of performance achieved by a person at a particular time. There is no way of knowing from a test score alone what caused that performance. Popular interpretations, however, continue to make the mistake of assuming that the cause of poor performance can be inferred from the score alone. This mistake can lead to several types of misleading conclusions including, for example, the erroneous conclusion that test scores can be equated with innate capacity or the erroneous conclusion that the educational quality of a state is below par solely because the average SAT score is low for that state.

Neither the general public nor policymakers has much awareness about what is on a typical standardized test, what it measures, or what factors influence the results, but they nonetheless often have great faith in test scores as indicators of school quality. For example, it frequently comes as a great surprise, even to otherwise knowledgeable policymakers, to learn

that the same form of an achievement test is administered for several years in a row to students within a district. When told this fact, the response is likely to be one of disbelief: "You don't mean that the same questions are asked each year do you?"

The "Lake Wobegon" effect, that is the tendency for most states and districts to report scores that are above the national average, is attributable, in part, to this reuse of the same test form year after year. Once a policymakers who ask about reuse of the same questions are convinced that, indeed, the same questions are asked year after year, they are quick to see the repeated use of a test form as a likely cause of the Lake Wobegon effect.

But why has something as obvious as the Lake Wobegon effect only recently come to our attention? The answer to this question may help explain another part of the general problem of test misuse.

Unfortunately, test results are not dependent only upon the questions that appear on the paper and on what students know. Rather, the results depend on a host of specific contextual factors of test use and administration. Some of these factors are obvious, albeit still often ignored. The dependence of state means on the SAT on the percentage of high school seniors within a state is an example of such an obvious, but often ignored, factor.

Other potentially influential factors are more subtle. The Lake Wobegon effect, for example, is largely the result of changes in the stakes that are attached to test results. Raising the stakes attached to the results for teachers and school

administrators increased the incentives to get scores up. That changed to context of testing and, in many cases, led to inflated scores. Certainly, the change in stakes has complicated interpretations of scores.

Students at school A may spend several weeks of class time on highly specific test preparation using commercially available practice materials keyed to the particular test that is used by the state or district. Students at school B, on the other hand, may not do any special preparation for the test. One could hardly expect the typical reader of a newspaper article comparing the scores for the two schools to know of this difference or understand its implications for interpreting the results. But such information is relevant for any judgment about the likely generalizability of the scores to broader content domains that the tests are supposed to sample.

It is unreasonable to expect policymakers, reporters, or the general public to become testing experts. In many other areas of concern to policymakers and the public, including, for example, health care, the environment, and the economy, the technical details of quantitative indicators remain largely obscure to the public and left in the hands of specialists. It is, after all, the job of specialists to take care of the technical details and provide overall results in a form that can be understood by the wider community.

Of course, technical experts are rarely of one mind. Testing experts disagree about the value of particular uses of tests and about the interpretations that can reasonably be

supported by test results, just as experts in medical research disagree about the degree of risk associated with drinking coffee and economists disagree about the governmental actions that should be taken based on recent results from leading economic indicators. As is true of other areas of specialization, there will always be cases where one testing specialist's misuse or misinterpretation is another's appropriate, or even highly recommended, use or interpretation.

Professional journals provide one forum for dealing with disagreements. But there are others as well, particularly where public policy is involved, including, for example, investigative reporting, legislative hearings, and judicial proceedings. All of these mechanisms have been used in the case of questions of test misuse. Resolving a particular issue by one of these mechanisms does not deal with the more systemic problems of test misuse, however. Furthermore, in many of the more prevalent examples there is a broad professional consensus that would oppose the misuse. Why is that opposition seemingly so ineffective in those instances?

The release of the first state-by-state results for the National Assessment of Educational Progress (NAEP) last June provides a recent example where there was not only a broad consensus but a major effort to prevent the use of the results to make misleading comparisons among states.[2] Thirty five states, two territories, and the District of Columbia participated in the first trial state assessment with the NAEP eighth grade mathematics assessment. There was a concerted effort to

encourage proper interpretations and discourage improper ones on the part of the National Assessment Governing Board, which has policy oversight of NAEP; the sponsoring governmental agency, the National Center of Educational Statistics (NCES); and the primary contractor, Educational Testing Service (ETS). In particular, the press was told that it was not appropriate to simply report the rank order of state results because many of the between-state differences in means were much too small to be statistically reliable.

NCES and ETS developed several alternative ways of reporting the results to emphasize the margin of error in each mean and to illustrate the large number of comparisons of state means where the differences were too small to be statistically reliable. Although the graphical presentation that was heavily relied on to make the point was quite imaginative and had great appeal to some of the more technically oriented audiences, it was ignored by the press because they considered it far too complex. Consequently, the seemingly more understandable simple rank order was reported in most news accounts of the results.

The point of this example is not to point the finger at either the press for failing to understand the more complex report or for believing that their readers would not understand the complexities. Nor is it to blame NCES or its contractor. On the contrary, the latter are to be complimented for seriously attending to the issue of potential misuse and for seeking ways of preventing it. Moreover, as was previously stated, it is unreasonable to expect the press to become testing experts. The

example, does suggest, however, that there is a need for more systematic and continuing effort on the part of test specialists to deal with the problems of test misuse and misinterpretation.

The effort to address issues of test misuse and misinterpretation needs to be at least twofold. First, a professional consensus needs to be sought that clearly identifies proper uses and justifiable interpretations as well as major prevalent misuses and misinterpretations. As previously indicated, there, of course, would be uses and interpretations about which there is not a clear professional consensus, but there eliminating those where there is a lack of agreement would still leave many important misuses for consideration.

Second, the business of communicating to a wider audience demands much greater attention than it has typically received. The ways in which the press, policymakers, and the general public interpret various presentations of test results is itself an area worthy of serious investigation. Market research is conducted prior to the introduction of many new products. Similar techniques could provide the information needed to judge the likely effectiveness of efforts to reduce test misuse and misinterpretation.

September 10, 1991

Footnotes.

1. Donlon, T. F. (1984). The Scholastic Aptitude Test. In T. F. Donlon (Ed.), The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests. New York: College Entrance Examination Board, p. 37.
2. Mullis, I. V. S., Dossey, J. A., Owen, E. H., & Phillips, G. W. (1991). The State of Mathematics Achievement: Executive Summary. Report No. 21-ST-03. Princeton, NJ: Educational Testing Service.